# Using syntax features and document discourse for relation extraction on PharmGKB and CTD

Schneider, Gerold ; Clematide, Simon ; Grigonyte, Gintare ; Rinaldi, Fabio

Abstract: We present an approach to the extraction of relations between pharmacogenomics entities like drugs, genes and diseases which is based on syntax and on discourse. Particularly, discourse has not been studied widely for improving Text Mining. We learn syntactic features semi-automatically from lean document-level annotation. We show how a simple Maximum Entropy based machine learning approach helps to estimate the relevance of candidate relations based on dependency-based features found in the syntactic path connecting the involved entities. Maximum Entropy based relevance estimation of candidate pairs conditioned on syntactic features improves relation ranking by 68% relative increase measured by AUCiP/R and by 60% for TAP-k (k=10). We also show that automatically recognizing document-level discourse characteristics to expand and filter acronyms improves term recognition and interaction detection by 12% relative, measured by AUCiP/R and by TAP-k (k=10). Our pilot study uses PharmGKB and CTD as resources.

DOI: https://doi.org/10.5167/uzh-64476

# Using syntax features and document discourse for relation extraction on PharmGKB and CTD

**Gerold Schneider, Simon Clematide, Gintarė Grigonytė, Fabio Rinaldi**
Institute of Computational Linguistics
University of Zurich
{gschneid,siclemat,gintare,rinaldi}@cl.uzh.ch

## Abstract

We present an approach to the extraction of relations between pharmacogenomics entities like drugs, genes and diseases which is based on syntax and on discourse. Particularly, discourse has not been studied widely for improving Text Mining. We learn syntactic features semi-automatically from lean document-level annotation. We show how a simple Maximum-Entropy based machine learning approach helps to estimate the relevance of candidate relations based on dependency-based features found in the syntactic path connecting the involved entities. Maximum Entropy based relevance estimation of candidate pairs conditioned on syntactic features improves relation ranking by 68% relative increase measured by AUCiP/R and by 60% for TAP-k (k=10). We also show that automatically recognizing document-level discourse characteristics to expand and filter acronyms improves term recognition and interaction detection by 12% relative, measured by AUCiP/R and by TAP-k (k=10). Our pilot study uses PharmGKB and CTD as resources.

## 1 Introduction

Pharmacogenomics and toxicogenomics study the relationships between drugs/chemicals, genes, and diseases, in particular in relation to specific individual mutations, which can affect the reactions to drugs and the susceptibility to diseases. Important databases that aim at providing a reference repository for such information are PharmGKB (Sangkuhl et al., 2008) and CTD (Wiegers et al., 2009). The information contained in PharmGKB and CTD is obtained from a combination of submitted experimental results and literature curation.

In this paper we describe research conducted by the OntoGene group within the scope of the SASEBio project (Semi-Automated Semantic Enrichment of the Biomedical Literature[1]), which aims at producing efficient Text Mining tools for the support of biomedical literature curation in realistic settings. We use the PharmGKB and CTD resources, which are large but have only lean document-level annotation: for each document, the IDs of relevant terms are given, but term occurrences or interaction evidence are not annotated.

## 2 Method

Our method for the extraction of interactions combines linguistic approaches, in particular syntactic analysis and discourse features. While many Text Mining tools in the biomedical and pharmacogenomics domain profit from syntactic features, discourse features have not been investigated and used widely yet.

### 2.1 Syntax-based approach

Approaches to the identification of entity interactions based on syntax are quite common. For example, (Fundel et al., 2007) describe a large-scale relation mining application using the Stanford Lexicalized Parser. Syntactic approaches can be further enhanced using machine learning methods, by extracting meaningful features from the dependency parse trees (e.g. (Erkan et al., 2007; Kim et al., 2008)).

We have parsed all sentences in the PharmGKB and in the CTD corpus with a dependency parser

[1] http://www.sasebio.org/

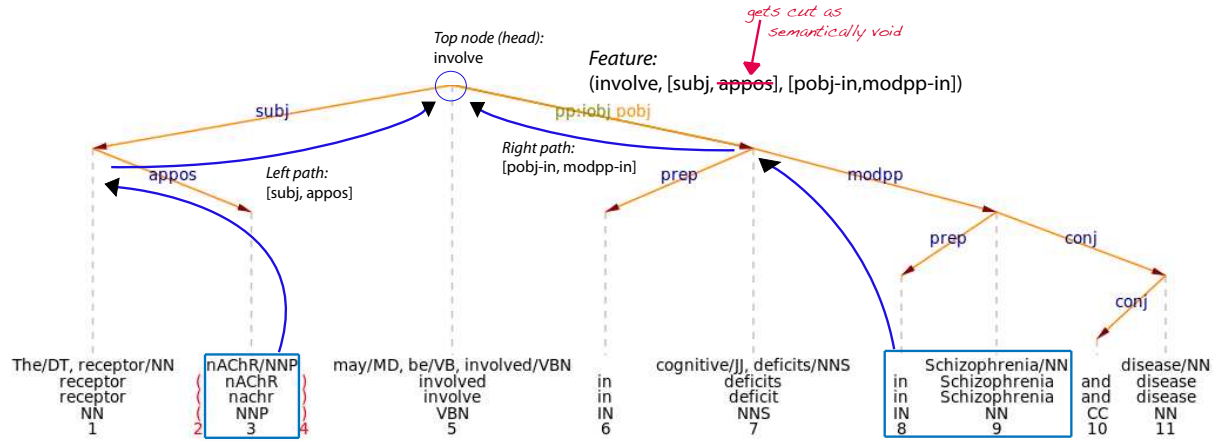


Figure 1: Simplified internal syntactic representation of the sentence "The **neuronal nicotinic acetylcholine receptor alpha7 (nAChR alpha7)** may be involved in cognitive deficits in Schizophrenia and Alzheimer's disease." from PubMed abstract 15695160. The curved arrows and dark red notes are aimed at illustrating the path features.

(Schneider, 2008). Lingpipe[2] is used for token and sentence segmentation. Term recognition is done by a dictionary-based tool which delivers annotated document spans (terms) associated to a set of identifiers (concepts) from domain term databases.

All entities that appear in the same sentence are potentially interacting, so we record the syntactic path that connects them as *candidate path*. A sample path is provided in Figure 1. If the gold standard states that both entities really interact in the document, then we mark the path that connects them as *relevant path*. The assumption that connecting paths between relevant entities are relevant allows us to use a weakly supervised approach, learning syntactic features from resources with lean, document-level annotation. The calculation of the number of *relevant paths* divided by the number of *candidate paths* gives us the Maximum-Likelihood probability that a path is relevant:

$$p(relevant) = \frac{freq(relevant\ path)}{freq(candidate\ path)}$$

The most frequent path types in the training set are given in Table 1. The third line, where the head word is *effect*, for example, has a modification by an of-PP to one of the entities in the relation, and a nested on-PP and of-PP modification. It covers patterns like *the effect of X on the increase of Y* or *no effect of X on the development of Y*, where X and Y are domain entities like drug, disease and protein.

We can use $p(relevant)$ directly during the application phase. Such a direct application, however, suffers from sparse data problems. We address this limitation by using half-paths (Section 3.1) and relevance probabilities computed by a Maximum-Entropy classifier (Section 3.2).

Similar approaches using PharmGKB as weakly supervised resource have been described in (Rinaldi et al., 2012) and in (Buyko et al., 2012). The latter also uses a feature-based classifier approach. Our experiment here differs in our explicit use of deep-linguistic resources like discourse (Section 2.3) and low-content or transparent words (Meyers et al., 1998), to avoid data sparseness, as follows: the relations for appositions, conjunctions and hyphens are cut from the path feature and parts of trees which are headed by a transparent word are cut. A transparent word is a word that does not affect the meaning of a sentence fundamentally if it is left out. For example, if *drug A affects groups of patients* then the sentence *drug A affects patients*, which does not contain the transparent word *group*, has a very similar meaning. We have learnt transparent words using the frequency-based approach of (Schneider et al., 2009): words that occur particularly often inside paths are regarded as transparent. The Genia corpus delivers over 300 transparent word types.

## 2.2 Maximum Entropy based estimation of path relevancy

In order to automatically estimate whether a syntactic path between two entities expresses a relevant re-

[2] http://alias-i.com/lingpipe/

| p(relevant) | Head | Path1 | Path2 | TP | Count |
|---|---|---|---|---|---|
| 13.62% | associate | subj | pobj-with | 53 | 389 |
| 17.82% | associate | subj modpp-in | pobj-with | 31 | 174 |
| 18.92% | effect | modpp-of | modpp-on modpp-of | 21 | 111 |
| 20.65% | association | modpp-of | modpp-with | 19 | 92 |
| 6.29% | be | obj modpp-of | subj | 19 | 302 |
| 17.82% | metabolize | pobj-by | subj | 18 | 101 |
| 29.63% | inhibit | pobj-by | subj | 16 | 54 |
| 23.81% | cause | subj modpp-in | obj | 15 | 63 |
| 100.00% | analyze | subj modpp-in | pobj-in modpart pobj-with | 14 | 14 |

Table 1: Some of the most frequent path types in the PharmGKB training set

lation we conducted a large-scale experiment with all curated relations from the CTD knowledge base where evidence from a PubMed article is mentioned. PubMed articles with more than 12 curated relations were omitted because it is unlikely that this amount of relations can be extracted by text mining methods from the very limited amount of text available in the abstracts. Our CTD corpus contains about 24,000 PubMed abstracts with about 72,000 relations in total. Test data (10%) and training data (90%) were sampled by stratification on the number of relations per article.

We use the Maximum Entropy classification tool[3] megam (Daumé, 2004) to learn the probability of a candidate path to be a relevant path, as described in section 2.1. Different sets of features derived from the candidate paths were used to build the conditional Maximum Entropy models for predicting the class probabilities.

In the result section, we present the result of four models: the baseline **B**, where the types of the entities are the only features we condition on. Our model **L** adds the following features to model **B** (complex features are noted between angle brackets): top head lemma,<entity1 type, top head lemma, entity2 type>, the unigrams of head lemmas from the paths, the bigrams of head lemmas from the paths. With model **L** we try to examine the contribution of syntactic heads for path relevancy estimation. Our model **D** adds the following features to model **B**: top head lemma, <entity1 type, head dependency to entity 1, top head lemma, head dependency to entity 2, entity2 type>, bigrams of head lemmas from the paths including the dependency label <head1,dependency,head2>. With model **D** we want to measure the contribution of syntactic dependency labels for relevancy estimation. Our model **DL** combines all features from **L** and **D**. For all models a threshold of 6 is applied to unigram features and a threshold of 3 to all others.

Our Maximum Entropy models compute the class probability of a single path, i.e. a mention of two entities in a single sentence. In order to compute the relevance score of a relation candidate for an entire abstract we take the mean of all probabilities from its path candidates. This relation score is then used for the ranking of all relation candidates.

### 2.3 Linguistic Discourse

Discourse investigates "a unit of language larger than a sentence and which is firmly rooted in a specific context" (Martin and Ringham, 2000, 51). Discourse is a broad area of linguistics, partly overlaps with pragmatics and includes a wide range of aspects, for example anaphora resolution, text genre studies, cohesion, felicity, and community-wide background knowledge. There are obvious ways in which discourse can help Text Mining. As salience of terms and frequency are closely related, the most frequently mentioned terms in a document are good interaction candidates and create a high baseline for protein-protein interaction approaches as we discuss in (Rinaldi et al., 2010). We investigate two aspects that are particularly relevant for relation detection in the biomedical domain.

First, many relations span several sentences, and if the two interacting entities are not in the same sentence, syntactic approaches thus fail. In PharmGKB, a third of all interaction pairs do not occur in the same sentence. Surface-based approaches, weighted

[3]For the training we used the binomial mode of megam which optimizes on the class probabilities, and we allowed for 200 iterations of feature weight optimizations. No bias feature was used because the skewed distribution of classes, i.e. very few items of class 1, and the large training set lead to errors when the bias feature was active.

| Method | Docs | TP | FP | FN | AUCiP/R | P | R |
|---|---|---|---|---|---|---|---|
| syn | 43 | 36 | 149 | 116 | 0.215 | 0.307 | 0.286 |
| syn.1side | 64 | 68 | 345 | 164 | 0.248 | 0.260 | 0.351 |
| syn.1side+appos | 65 | 71 | 351 | 163 | 0.256 | 0.266 | 0.361 |
| syn+cooc | 73 | 116 | 1044 | 151 | 0.277 | 0.143 | 0.477 |
| syn+cooc2 | 72 | 158 | 2337 | 106 | 0.279 | 0.094 | 0.616 |
| syn+cooc2w | 72 | 165 | 2685 | 99 | 0.286 | 0.091 | 0.650 |
| syn+cooc2wf | 72 | 167 | 3783 | 97 | 0.286 | 0.073 | 0.661 |

Table 2: Evaluation of 75 manually annotated PharmGKB documents. The first column gives the approach used. The second column reports the number of documents with at least one response hit. The third to the fifth column give true positives (TP), false positives (FP) and false negatives (FN). The sixth column contains the macro averaged AUCiP/R. The seventh column reports macro precision, the eighth macro recall.

by distance, increase recall, as we discuss in Section 3.1.

Second, term detection integrating document-level information can improve the results of a dictionary-based term-recognition approach. We profit from the whole document both to increase recall and precision of term recognition, as we describe in the following, and give results in Sections 3.1 and 3.2. (Schwartz and Hearst, 2003) introduce an algorithm for detecting acronyms in brackets. Our approaches go beyond this by using a more general syntactic relation, and by profiting from concept references.

**Expanding introduced acronyms** Abbreviations are often introduced inside a document with the apposition relation. Figure 1 shows an example of an apposition relation connecting a full form (*neuronal nicotinic acetylcholine receptor alpha7*) to an acronym (*nAChR alpha7*)[4]. Short acronyms are often highly ambiguous or the correct concept reference cannot be found. We add the expansion to all acronyms that are introduced in a document, if their concepts differ. This step increases recall at the cost of precision.

(1) *The current studies were designed to examine if quinone intermediates are involved in the toxicity of hepatotoxic halobenzenes, bromobenzene (BB) and 1,2,4-trichlorobenzene (1,2,4-TCB).* (CTD, pubmed 10092053)

In sentence (1) the acronym *BB* is given a gene concept by the term recognizer, while it is an acronym of the chemical substance *bromobenzene*, to which *BB* is connected via a syntactic apposition relation. All 5 occurrences of *BB* in the document are thus given the chemical concept of *bromobenzene*.

[4]The graph simplifies the sentence for reasons of space.

**Filtering acronyms without expansion candidates** We refer to the process of mapping an acronym to its long form as expansion. Those concepts of short acronyms which do not have promising expansion candidates in the document are filtered out. This step increases precision at the cost of recall. We consider short words (up to 4 characters) as acronyms. We check the list of terms found in the document against the list of variants of terms in the reference terminology. For instance, in the PubMed citation 12932788, our pipeline finds the following 15 term candidates: *LXRalpha, cholesterol, bile acid, glucose, LBD, retinoic acid receptor gamma, RARgamma, all-trans retinoic acid, 22(R)-hydroxycholesterol, benzenesulfonamide, T0901317, diphenyl, phenyl-acetic acid, GW3965, toa*. Two of these terms are considered acronyms and are checked against the reference terminology: *LBD (MESH:D020192)* and *toa (CTD:100008541)*.

*LBD* is referring to 'LXRalpha ligand-binding domain', but it was recognized as the disease term 'Lewy Body Dementia'. We check if any other variant listed under the concept *MESH:D020192* occurs in the text. In the case of *LBD* it is not, therefore the term *LDB* referring to the disease is removed.

Concerning *toa* the term recognizer maps it to gene ID *CTD:100008541* due to our aggressive candidate generation, but it actually refers to the sequence 'to a' in the text. No other variants of the concept *CTD:100008541* can be found in the text and therefore *toa* is also discarded.

| DL: Dependency + Lemma | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Appos** | **Termfilter** | **Transparent** | Docs | TP | FP | FN | AUCiP/R | TAP-10 | P | R | F |
| - | - | - | 2233 | 2525 | 30318 | 4468 | 0.27831 | 0.2168 | 0.09842 | 0.41289 | 0.14271 |
| + | - | - | 2239 | 2845 | 34490 | 4165 | 0.31292 | 0.2428 | 0.09938 | 0.45634 | 0.14694 |
| + | + | - | 2182 | 2368 | 23394 | 4501 | 0.27316 | 0.1811 | 0.11832 | 0.39170 | 0.16136 |
| + | + | + | 2182 | 2368 | 23394 | 4501 | 0.27459 | 0.1819 | 0.11832 | 0.39170 | 0.16136 |
| **D: Dependency** | | | | | | | | | | | |
| **Appos** | **Termfilter** | **Transparent** | Docs | TP | FP | FN | AUCiP/R | TAP-10 | P | R | F |
| - | - | - | 2233 | 2525 | 30318 | 4468 | 0.28693 | 0.2223 | 0.09842 | 0.41289 | 0.14271 |
| + | - | - | 2239 | 2845 | 34490 | 4165 | 0.30756 | 0.2391 | 0.09938 | 0.45634 | 0.14694 |
| + | + | - | 2182 | 2368 | 23394 | 4501 | 0.27664 | 0.1835 | 0.11832 | 0.39170 | 0.16136 |
| + | + | + | 2182 | 2368 | 23394 | 4501 | 0.28024 | 0.1854 | 0.11832 | 0.39170 | 0.16136 |
| **L: Lemma** | | | | | | | | | | | |
| **Appos** | **Termfilter** | **Transparent** | Docs | TP | FP | FN | AUCiP/R | TAP-10 | P | R | F |
| - | - | - | 2233 | 2525 | 30318 | 4468 | 0.27992 | 0.2180 | 0.09842 | 0.41289 | 0.14271 |
| + | - | - | 2239 | 2845 | 34490 | 4165 | 0.30840 | 0.2401 | 0.09938 | 0.45634 | 0.14694 |
| + | + | - | 2182 | 2368 | 23394 | 4501 | 0.27244 | 0.1806 | 0.11832 | 0.39170 | 0.16136 |
| + | + | + | 2182 | 2368 | 23394 | 4501 | 0.27384 | 0.1814 | 0.11832 | 0.39170 | 0.16136 |
| **B: Baseline** | | | | | | | | | | | |
| **Appos** | **Termfilter** | **Transparent** | Docs | TP | FP | FN | AUCiP/R | TAP-10 | P | R | F |
| - | - | - | 2233 | 2525 | 30318 | 4468 | 0.16599 | 0.1351 | 0.09842 | 0.41289 | 0.14271 |
| + | - | - | 2239 | 2845 | 34490 | 4165 | 0.17206 | 0.1374 | 0.09938 | 0.45634 | 0.14694 |
| + | + | - | 2182 | 2368 | 23394 | 4501 | 0.16514 | 0.1071 | 0.11832 | 0.39170 | 0.16136 |
| + | + | + | 2182 | 2368 | 23394 | 4501 | 0.18360 | 0.1188 | 0.11832 | 0.39170 | 0.16136 |

Table 3: Evaluation of the CTD corpus. The first 3 columns give the approach used. The other columns are analogous to Table 2, with the addition of a column for TAP-10, and one for F-score.

## 3 Results

We have applied our approach to a manually verified test set from PharmGKB and to the entire CTD.

### 3.1 Results from PharmGKB

Evaluation results[5] on PharmGKB are given in Table 2. The method **syn** is purely our syntactic method, as described in Section 2.1. The method **syn.1side** uses half-path features as a backoff. If either the left or the right side from a term to the top node match to a decision from the gold standard, the decision is reported. The method **syn.1side+appos** additionally recognizes acronyms that were introduced by a syntactic apposition relation. The relatively low recall of syntactic methods can be increased by including sentence-coocurrence, which the method **syn+cooc** does. We can see on the one hand that recall increases at the cost of precision, on the other hand that it is still below 50%, which indicates that many interactions are expressed across several sentences. The method **syn+cooc2** extends the sentence-coocurrence score to including the neighbouring sentence. The increase in recall indicates that context of more than one sentence is often necessary. The method **syn+cooc2w** weighs the sentence-coocurrence score by distance, giving higher scores to entities that appear closer. The method **syn+cooc2wf** is identical but does not use a score threshold, thus returning all results, which increases recall and reduces precision. It aims to give an upper bound on recall.

The evaluation results of (Buyko et al., 2012) are not comparable to the results presented in Table 2. They evaluate on a specifically crafted subcorpus where both participating entities have to appear in a single sentence. Additionally they cover only relations between entities of different types, i.e. gene-disease, gene-drug, drug-disease.

### 3.2 Results from CTD with Maximum Entropy

The evaluation results from CTD for our approach described in Section 2.2 are shown in Table 3. We have also tested the expansion of introduced acronyms (**appos**) and the filtering of unexpanded acronyms (**termfilter**) as suggested in Section 2.3.

Our experiments focus on evaluation metrics reflecting the quality of the ranking of candidate relations: AUCiP/R and TAP-k. AUCiP/R measures the

[5] We use the BioCreative scorer from `http://www.biocreative.org/tasks/biocreative-ii5/biocreative-ii5-evaluation/` with default settings, which ignores null documents

area under the interpolated Precision/Recall curve. TAP-k (Threshold Average Precision, (Carroll et al., 2010)) averages precision for the results above a given error threshold k. These measures directly relate to the expected user's benefit in a curation scenario. For this evaluation where we expect no more than 12 true positive relations we set $k = 10$. Precision, recall and F-score are also given to show that **Termfilter** has best F-score. If we added cross-sentential term-cooccurrence features, recall would probably increase in a similar fashion as for PharmGKB in section 3.1, but that was not the goal of this experiment.

The dependency model (**D**), the lemma model (**L**), and the combined model (**DL**) perform substantially better than the baseline (**B**), improving relation ranking by 68%. **Appos** shows relative improvements for all evaluations metrics, **DL** improves by 12%. **Termfilter** leads to better precision and better F-score, but AUCiP/R and TAP-k suffer. Cutting **transparent** words leads to a marginally higher performance, further investigations are needed here.

## 4 Conclusions

We have presented two approaches to the extraction of relations between pharmacogenomics entities, based on learning syntactic features semi-automatically from lean document-level annotation. We have shown how a simple Maximum-Entropy based machine learning approach helps to estimate the relevance of candidate relations when using dependency-based features found in the syntactic path connecting the involved entities. Maximum-Entropy based relevance estimation of candidate pairs conditioned on syntactic features improves relation ranking by 68% relative increase measured by AUCiP/R and by 60% for TAP-10, with respect to a baseline method that conditions solely on the distribution of the type of entities.

We have suggested and implemented methods which profit from the document as a discourse entity. Discourse has hardly been investigated for improving Text Mining before. We show that our method, which expands and filters acronyms, improves term recognition and interaction detection by 12% in terms of AUCiP/R and TAP-10. Our research on discourse also shows that document-level discourse characteristics improve term recognition and Text Mining. As future research, we plan to integrate syntactic evidence and surface-based approaches for relation mining into annotation tools for the support of biomedical database curators.

## 5 Acknowledgements

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel.

## References

Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2012. Extraction of pharmacogenetic and pharmacogenomic relations – a case study using pharmgkb. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pages 376–387, Hawaii.

Hyrum D Carroll, Maricel G Kann, Sergey L Sheetlin, and John L Spouge. 2010. Threshold Average Precision ( TAP-k ): A Measure of Retrieval Designed for Bioinformatics. *Methods*, pages 1–8.

Hal III Daumé. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at http://pub.hal3.name#daume04cg-bfgs, implementation available at http://hal3.name/megam/.

G. Erkan, A. Ozgur, and D. R. Radev. 2007. Extracting interacting protein pairs and evidence sentences by using dependency parsing and machine learning techniques. In *Proceedings of BioCreAtIvE 2*.

K. Fundel, R. Küffner, and R. Zimmer. 2007. RelEx – relation extraction extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

S. Kim, J. Yoon, and J. Yang. 2008. Kernel approaches for genic interaction extraction. *Bioinformatics*, 9:10.

Bronwen Martin and Felizitas Ringham. 2000. *Dictionary of Semiotics*. Cassell, New York.

Adam Meyers, Catherine Macleod, Roman Yangarber, Ralph Grishman, Leslie Barrett, and Ruth Reeves. 1998. Using NOMLEX to produce nominalization patterns for information extraction. In *Coling-ACL98 workshop Proceedings: the Computational Treatment of Nominals*, Montreal, Canada.

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Thérèse Vachon, and Martin Romacker. 2010. Ontogene in biocreative ii.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 3:472–480.

Fabio Rinaldi, Gerold Schneider, and Simon Clematide. 2012. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics*.

Katrin Sangkuhl, Dorit S. Berlin, Russ B. Altman, and Teri E. Klein. 2008. PharmGKB: Understanding the effects of individual genetic variants. *Drug Metabolism Reviews*, 40(4):539–551. PMID: 18949600.

Gerold Schneider, Kaarel Kaljurand, and Fabio Rinaldi. 2009. Detecting Protein/Protein Interactions using a parser and linguistic resources. In *CICLing 2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 406–417, Mexico City, Mexico. Springer LNC 5449.

Gerold Schneider. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

AS Schwartz and MA Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput*, pages 451–462.

T Wiegers, A Davis, KB Cohen, L Hirschman, and C Mattingly. 2009. Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (ctd). *BMC Bioinformatics*, 10(1):326.